

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-105500

(43) 公開日 平成10年(1998) 4月24日

(51) Int.Cl.⁶
G 0 6 F 13/00
H 0 4 L 12/40
12/28

識別記号
3 5 7

F I
G 0 6 F 13/00
H 0 4 L 11/00
11/20
3 5 7 Z
3 2 1
C

審査請求 未請求 請求項の数 8 O L (全 9 頁)

(21) 出願番号 特願平9-243266

(22) 出願日 平成9年(1997) 9月9日

(31) 優先権主張番号 0 8 / 7 1 1 1 8 9

(32) 優先日 1996年9月9日

(33) 優先権主張国 米国 (U S)

(71) 出願人 596092698

ルーセント テクノロジーズ インコーポ
レーテッド

アメリカ合衆国. 07974-0636 ニュージ
ヤージー, マレイ ヒル, マウンテン ア
ヴェニュー 600

(72) 発明者 ヴラディミール ネプスティル

アメリカ合衆国 80303 コロラド, ポー
ルダール, クアラ ドライブ 4905

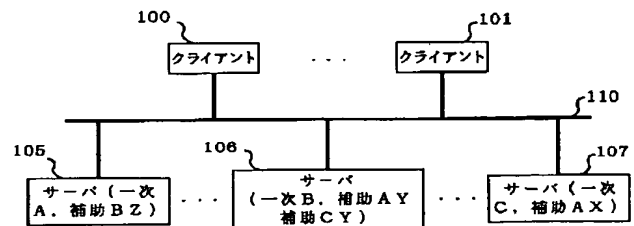
(74) 代理人 弁理士 岡部 正夫 (外11名)

(54) 【発明の名称】 ネットワーク・サーバの動的再構成

(57) 【要約】

【課題】 本発明は、インターネットやイントラネットのアーキテクチャ等のクライアント-サーバシステムのアーキテクチャに関し、特に、サーバの処理負荷を分散する技術を提供する。

【解決手段】 本発明のクライアント-サーバシステムは、クライアント要求を処理する複数のサーバからなり、あるサーバ上の処理負荷が過剰でない間は、全クライアント要求を処理し、該少なくとも1つの第1のサーバ上の処理負荷が過剰になった場合には、ある特定のクライアント要求のみを処理して他のクライアント要求は他のサーバに対して自動的にリダイレクトし、そしてリダイレクト先のサーバはそのリダイレクションに自動的に応答して、リダイレクトされたクライアント要求を処理することを特徴とする。



【特許請求の範囲】

【請求項1】 クライアント・サーバ・システムであって、

クライアント要求を処理する複数のサーバからなり、
該複数のサーバのうちの少なくとも1つの第1のサーバは、第1の情報及び該第1の情報に関連する第2の情報を有し、該少なくとも1つのサーバ上の処理負荷が過剰でない間は、該第1の情報を必要とする該クライアント要求の部分と該第2の情報を必要とする該クライアント要求の部分とを処理し、該少なくとも1つの第1のサーバ上の処理負荷が過剰になったことに応答して、第2の情報を必要とする該クライアント要求の部分の処理することなく第1の情報を必要とする該クライアント要求の部分の処理し、処理をする少なくとも1つの第2のサーバに対して第2の情報を必要とする該クライアント要求の部分の自動的にリダイレクトするものであり、そして、

該複数のサーバのうちの該少なくとも1つの第2のサーバは、該第1および第2の情報のうちの第2の情報を有しており、該リダイレクションに自動的に応答して、該第2の情報を必要とする該クライアント要求の該リダイレクトされた部分を処理することを特徴とするクライアント・サーバ・システム。

【請求項2】 請求項1に記載のシステムにおいて、該少なくとも1つの第2のサーバは、該第1の情報を有することなく該第2の情報を有しており、該第1の情報を必要とするクライアント要求の部分の処理することなく、第2の情報を必要とする該クライアント要求のリダイレクトされた部分を処理することを特徴とするシステム。

【請求項3】 請求項1に記載のシステムにおいて、該少なくとも1つの第1のサーバ上の処理負荷が過剰でない間は、該第1の情報は該少なくとも1つの第1のサーバの該第2の情報を示す第1のページまたはオブジェクトを含んでおり、該少なくとも1つの第1のサーバ上の処理負荷が過剰になったことに応答して、該少なくとも1つの第2のサーバの該第2の情報へ示す第2のページまたはオブジェクトを含むことを特徴とするシステム。

【請求項4】 請求項1に記載のシステムにおいて、該少なくとも1つの第1のサーバ上の処理負荷が所定の限界値を超えていない間は、該少なくとも1つの第1のサーバは該クライアント要求の両方の部分を処理し、該少なくとも1つの第1のサーバ上の処理負荷が所定の限界値を超えたことに応答して、該第2の情報を必要とする該クライアント要求の部分の処理することなく該第1の情報を必要とする該クライアント要求の部分の処理し、該第2の情報を必要とする該クライアント要求の部分の自動的にリダイレクトすることを特徴とするシステム。

【請求項5】 請求項4に記載のシステムにおいて、

該所定の限界値は第1の所定の第1の限界値であり、該少なくとも1つの第1のサーバ上の処理負荷が所定の第2の限界値以下になったことに応答して、該少なくとも1つの第1のサーバは、該第2の情報を必要とする該クライアント要求の部分のリダイレクトすることを自動的に止め、該クライアント要求の両方の部分の処理を回復することを特徴とするシステム。

【請求項6】 請求項1に記載のシステムにおいて、

10 該少なくとも1つの第1のサーバはさらに、該第1の情報に関連付けられている第3の情報をさらに有し、少なくとも1つのサーバ上の処理負荷が所定の第1の限界値を超えていない間は、該第1の情報を必要とする該クライアント要求の部分と、該第2の情報を必要とする該クライアント要求の部分と、該第3の情報を必要とする該クライアント要求の部分とを処理し、該少なくとも1つの第1のサーバ上の処理負荷が該所定の第1の限界値を初めて超過したことに応答して、該第2の情報を必要とする該クライアント要求の部分の処理することなく該第3の情報を必要とする該クライアント要求の部分と第1の情報を必要とする該クライアント要求の部分とを処理し、該第2の情報を必要とする該クライアント要求の部分の、処理をする該少なくとも1つの第2のサーバへ自動的にリダイレクトし、そして該少なくとも1つの第1のサーバ上の処理負荷が該所定の第1の限界値を二度目に超えたことに応答して、該第2の情報を必要とする該クライアント要求の部分および該第3の情報を必要とする該クライアント要求の部分の処理することなく該第1の情報を必要とする該クライアント要求の部分の処理

30 し、該第3の情報を必要とする該クライアント要求の部分の該複数のサーバの処理する少なくとも1つの第3のサーバへ自動的にリダイレクトし、
該複数のサーバのうちの該少なくとも1つの第3のサーバは該第1、第2、および第3の情報のうちの第3の情報を有し、該少なくとも1つの第1のサーバ上の処理負荷が該所定の第1の限界値を二度目に超えた時に、該第3の情報を必要とする該クライアント要求の該リダイレクトされた部分を処理し、そして、

40 該少なくとも1つの第1のサーバ上の処理負荷が所定の第2の限界値以下に最初に落ちたことに応答して、該少なくとも1つの第1のサーバは更にリダイレクションを自動的に止めて、該第2及び第3の情報を必要とする該クライアント要求の部分の処理を回復し、そして該少なくとも1つの第1のサーバ上の処理負荷が所定の第2の限界値以下に二度目に落ちたことに応答して、リダイレクションを自動的に止めて、該第2および第3の情報のうちの1つを必要とする該クライアント要求の部分の処理を回復することを特徴とするシステム。

50 【請求項7】 クライアント要求を処理する複数のサーバからなるクライアント-サーバ システムを操作する

方法であって、該方法が、
 該複数のサーバのうちの少なくとも1つの第1のサーバ
 において、該少なくとも1つの第1のサーバ上の処理負
 荷が過剰でない間は、該第1の情報を必要とする該ク
 ライアント要求の部分と該第1の情報に関連付けられて
 いる第2の情報を必要とする該クライアント要求の部分
 の両方を処理する段階からなり、該少なくとも1つの第
 1のサーバは該第1の情報と該第2の情報の双方を有し
 ており、該システムは更に、
 該少なくとも1つの第1のサーバ上の処理負荷が過剰に
 なってきたことに応答して、該少なくとも1つの第1の
 サーバの中の第2の情報を必要とする該クライアント要
 求の部分の処理することなく該第1の情報を必要とする
 該クライアント要求の部分の処理する段階と、
 該少なくとも1つの第1のサーバ上の処理負荷が過剰に
 なってきたことに応答して、
 該第2の情報を必要とする該クライアント要求の部分
 を複数のサーバのうちの該少なくとも1つの第1のサーバ
 へ自動的にリダイレクトする段階とからなり、該少なく
 とも1つの第2のサーバは第2の情報を有しており、該
 方法は更に、
 該リダイレクションに応答して、該少なくとも1つの第
 2のサーバにおいて該第2の情報を必要とする該ク
 ライアント要求のリダイレクトされた部分を自動的に処理
 する段階からなることを特徴とする方法。

【請求項8】 請求項7に記載の方法においてさらに、
 該少なくとも1つの第1のサーバ上の処理負荷が所定の
 レベル以下に落ちたことに応答して、該第2の情報を必
 要とする該クライアント要求の該部分をリダイレクトす
 ることを自動的に止める段階と、
 該少なくとも1つの第1のサーバ上の処理負荷が該所定
 のレベル以下に落ちたことに応答して、該少なくとも1
 つのサーバの中の該クライアント要求の該部分の両方
 の処理を自動的に回復する段階とからなることを特徴と
 する方法。

【発明の詳細な説明】

【0001】

【発明の分野】 本発明は、インターネットやイントラネ
 ットのアーキテクチャなどの、情報ネットワーク・アー
 キテクチャに関する。

【0002】

【発明の背景】 インターネットなどの情報ネットワーク
 において、クライアントと呼ばれるユーザのコンピュー
 タが、サーバと呼ばれる情報プロバイダのコンピュータ
 から情報を要求し、そのサーバは要求された情報をその
 クライアントに対して供給する。インターネット上での
 情報の格納、検索、および転送のための事実上の標準
 (de-facto standard) であるワールド・ワイド・ウェ
 ブ (World Wide Web: WWW) において、情報はページ
 の形式で供給される。ページはテキスト、グラフィッ

ク、スクリプト、または他の形式で表現される情報のデ
 ィスプレイの全面面である。ページは1つまたはそれ以
 上のオブジェクトを含む。オブジェクトはユニホーム・
 リソース・ロケータ (Uniform Resource Locator: UR
 L) と呼ばれる、それ自身のネットワーク・アドレス
 (ユニークな単独のアドレスであることが好ましい) を
 有する情報要素である。例えば、ページはディスプレイ
 画面上で1つまたはそれ以上のテキスト・オブジェク
 ト、1つまたはそれ以上の画像オブジェクト、およびフ
 レーム・オブジェクトによって定義されるレイアウトの
 中に表示される1つまたはそれ以上のスクリプト・オブ
 ジェクトを含むことができる。

【0003】 通常、サーバは自分が提供する情報および
 サービスに対する入り口 (entry point) として役立つ
 メイン・ページを備えている。このページは通常、同じ
 サーバによってサービスされる他のページおよびオブジ
 ェクト (例えば、グラフィック画像、ビデオ/オーディ
 オ/テキスト・ファイル) などを示している。

【0004】 一般に、クライアントがサーバにアクセス
 する時、サービスはそのクライアントに対してメイン・
 ページを提供し、次にクライアントと対話してそのク
 ライアントに必要な追加の情報またはサービスを提供す
 る。サーバにアクセスするクライアントの数が増加する
 につれて、サーバの処理負荷が増加し、その性能は実質
 的に劣化する。従って、ユーザがサーバに対して要求を
 出してからその要求がサーバによって満足されるまでの
 間の遅延時間が増加することになる。

【0005】 サーバの過負荷を避けるために、一般に管
 理者はそのサーバを手動で再構成しなければならず、そ
 の対象のサーバ上の負荷を軽くするために、要求のいく
 つかを他のサーバに対してリダイレクトする。いくつか
 のサービス・プロバイダはサービスされる情報の複製を
 複数のサーバの中に記憶し、異なるサーバが異なる要求
 に対して、例えば、総当たり (round-robin) を基礎と
 してサービスする。これによって複数のサーバに要求の
 負荷を分散する。これにはいくつかの欠点がある。まず
 第1に、管理者の手動での介入は遅くて不十分であり、
 誤りを起こしやすく、そして即座には行われなことが
 多い。第2に、総当たりを基礎として要求にサービスす
 るために複数のサーバを使うことによって、要求が比較
 的少数である期間においてサーバの利用率が低下するこ
 とになり、従って、それは非効率的である。さらに、こ
 の方式ではすべてのサーバの情報が各サーバ上に複製さ
 れている必要があり、サーバは共通のデータに対する共
 通のキャッシュを利用することができない。

【0006】

【発明の概要】 本発明は、従来の技術のこれらの問題点
 および他の問題点を解決することを指向している。一般
 に、本発明によると、クライアント要求を処理するため
 に、一次サーバが使用する情報の一部分が1つまたはそ

れ以上の補助の待機サーバ上に複製され、クライアントのサービスに対する需要が増加すると、一次サーバ上の処理負荷が過剰になり、一次サーバは情報の複製された部分を必要とするクライアント要求の部分の処理を自動的に補助サーバに肩代わりさせる。サービスに対する需要が減少し、一次サーバが過小負荷状態になると、クライアント要求全体を一次サーバが自動的にサービスするように回復する。

【0007】本発明の利点は、その時点での処理負荷に基づいて、人間の介入なしに、負荷の分散および負荷の共有が自動的に発生することを含む。サーバ全体の中から1台のサーバまたは1グループのサーバだけが任意の時点において、情報の個々の部分（例えば、ページ、またはオブジェクト、ページまたはオブジェクトのグループ）にサービスし、それによって情報の効率的なキャッシングが可能となる。サービスに対するクライアントの需要が大幅に変動する場合であっても、比較的一様な応答時間がクライアントに対して提供される。さらに、待機サーバが一次サーバのクライアントにサービスしていない間、その処理パワーを他の処理作業のために使うことができる。例えば、サービスに対する需要のピークの時刻が一次サーバのクライアントとは異なっている他のクライアントにサービスし、それによってサーバを効率的に利用することができる。

【0008】本発明の第1の側面によれば、クライアント・サーバ・システムはクライアント要求を処理するために複数のサーバを含み、複数のサーバのうちの少なくとも1つの第1のサーバが、第1の情報を必要とするクライアント要求の部分および第2の情報を必要とするクライアント要求の部分処理するために、第1の情報およびその第1の情報に関連付けられた第2の情報を有している。その少なくとも1つのサーバ上の処理負荷が過剰でない間、例えば、所定の第1の限界値を超えていない間、その少なくとも1つの第1のサーバがクライアント要求の両方の部分を処理する。少なくとも1つの第1のサーバ上の処理負荷が過剰になったことに応答して、その少なくとも1つの第1のサーバは第2の情報を必要とするクライアント要求の部分処理せずに、第1の情報を必要とするクライアント要求の部分処理し、第2の情報を必要とするクライアント要求の部分処理するために、少なくとも1つの第2のサーバに対して自動的にリダイレクトする。複数のサーバのうちの少なくとも1つのサーバがその第2の情報を有し、第2の情報を必要とするクライアント要求のリダイレクトされた部分をそのリダイレクションに回答して自動的に処理する。好ましくは、少なくとも1つの第1のサーバ上の処理負荷が、例えば、所定の第2の限界値以下に落ちたことに応答して、その少なくとも1つの第1のサーバはその第2の負荷を必要とするクライアント要求の部分のリダイレクトすることを止め、そのクライアント要求の両方の部

分の処理を回復する。

【0009】本発明の第2の側面によれば、クライアント要求を処理するための複数のサーバを含むクライアント・サーバ・システムを動作させる方法は、次のステップを含む。複数のサーバのうちの少なくとも1つの第1のサーバ上の処理負荷が過剰でない間、その少なくとも1つの第1のサーバは第1の情報を必要とするクライアント要求の部分と、その第1の情報に関連付けられた第2の情報を必要とするクライアント要求の部分の両方を処理し、その少なくとも1つの第1のサーバは第1の情報と第2の情報の両方を持っている。少なくとも1つの第1のサーバ上の処理負荷が過剰になってきたことに応答して、その少なくとも1つのサーバは第2の情報を必要とするクライアント要求の部分処理せずに、第1の情報を必要とするクライアント要求の部分処理し、第2の情報を必要とするクライアント要求の部分処理、複数のサーバのうちの少なくとも1つの第2のサーバに対して自動的にリダイレクトする。そのリダイレクションに回答して、その少なくとも1つの第2のサーバは第2の情報を必要とするクライアント要求のリダイレクトされた部分を自動的に処理し、その少なくとも1つの第2のサーバは第2の情報を持っている。好適には、少なくとも1つのサーバ上の処理負荷が所定の限界値以下になった時、その少なくとも1つのサーバは第2の情報を必要とするクライアント要求の部分処理をリダイレクトすることを自動的に止め、クライアント要求の両方の部分の処理を回復する。

【0010】本発明のこれらおよび他の利点および特徴は図面と一緒に本発明の例示としての実施形態の次の記述から、より明らかになる。

【0011】

【発明の詳細な記述】図1は例示としての情報ネットワーク（この例においてはインターネット）を示している。それは複数のクライアント100乃至101および複数のサーバ105乃至107がインターネットのネットワーク構造110によって相互に接続されているものを含む。各サーバ105はプロセッサおよびメモリを備えているコンピュータであり、その中でプロセッサはそのメモリの中に格納されている制御プログラムを実行して、そのメモリの中に格納されているサービスおよびデータを提供する。各サーバ105乃至107はそれぞれ、情報A乃至Cのデータベースに対する一次サーバである。しかし、本発明によると、データベースに対する一次サーバである以外に、各サーバ105乃至107は1つまたはそれ以上の他のサーバのデータベースの一部に対する第2のサーバ、すなわち、サポートしているサーバでもある。図1の例において、サーバ105はサーバ106のデータベースBの部分BZに対する補助サーバであり、サーバ106はサーバ105のデータベースAの部分AYおよびサーバ107のデータベースCの

一部分CWに対する補助サーバであり、そしてサーバ107はサーバ105のデータベースAの一部分AXに対する補助サーバである。

【0012】図2は本発明を理解するのに重要であるサーバ105乃至107のメモリ205乃至207の内容をそれぞれ示している。メモリ205はデータベースA210と、データベースB220の部分BZ221の複製211とを格納している。メモリ206はデータベースB220と、データベースC230の部分CW231の複製222と、データベースA210の部分AY213の複製223とを格納している。そして、メモリ207はデータベースC230と、データベースA210の部分AY212の複製232とを格納している。

【0013】データベースA210はデータベースA210の中の他のページA2(243)乃至A4(245)に対するリンクを含む一次メイン・ページA1(240)を含む。さらに、ページA3(244)、A4(245)はそれぞれ他のページA5(246)およびA6(247)乃至A7(248)に対するリンクもそれぞれデータベースA210の中に含む。データベースA210はデータベース210の中のページA2(243)乃至A3(244)に対するリンクを含む第2のメイン・ページA1'(241)をさらに含むが、メモリ205の中の部分AX212のページA4(245)に対するリンクを含む代わりに、メモリ207の中の複製部分AX232の複製ページA4(245)に対するリンクを含む。さらに、データベースA210はデータベース210の中のページA2(243)に対するリンクを含む第3のメイン・ページA1''(242)をさらに含むが、メモリ205の中の部分AX212およびAY213のページA3(244)およびA4(245)に対するリンクを含む代わりに、メモリ206および207のそれぞれの中の複製部分AY223およびAX232の複製ページA3(244)およびA4(245)に対するリンクをそれぞれ含む。

【0014】データベースB220はオブジェクトB2(251)に対するリンクおよびデータベースB220の中の別のページB3(252)に対するリンクを含んでいるメイン・ページB1(250)を含む。B3(252)はデータベースB220の中の他のオブジェクトまたはページB4(258)乃至B5(259)に対するリンクをさらに含む一次ページである。データベースB220はメモリ206の中のデータベースB220の部分BZ221の中のページB4(258)乃至B5(259)に対するリンクを含む代わりに、メモリ205の中の複製部分BZ221の複製オブジェクトおよび/またはページB4(258)乃至B5(259)に対するリンクを含む第2のページB3'(253)をさらに含む。

【0015】データベースC230はデータベースC2

30の部分CW231の中のデータ・オブジェクトC262に対するリンクを含む、メイン・ページの一次フォーマット・オブジェクトC1(260)を含む。データベースC230はメモリ207の中のデータベースC230の部分CW231の中のデータ・オブジェクトC2(262)に対するリンクを含む代わりに、メモリ206の中の複製部分CW222の複製オブジェクトC2(262)に対するリンクを含む、メイン・ページの二次的なフォーマット・オブジェクトC1'(261)をさらに含む。

【0016】サーバ105乃至107は処理負荷が最も重くなる時点がそれぞれ異なると仮定される。例えば、サーバ105は夕方において最も頻繁に使われる可能性があり、サーバ106は週日において最も頻繁に使われる可能性があり、サーバ107は週末に使われることが最も多い可能性がある。最初に、すべてのサーバ105乃至107は普通に動作する。すなわち、サーバ105はデータベースA210からの情報に対するすべての要求にサービスし、サーバ106は最初にデータベースB220からの情報に対するすべての要求にサービスし、そしてサーバ107はデータベースC230からの情報に対するすべての要求にサービスする。それは従来の方法で、一次ページおよびオブジェクト240、252、260、またはこれらのいずれか一方を使って行われる。また普通のように、各サーバ105乃至107は、例えば、単位時間当たりにサービスされたアクセス(要求)の回数の形で、現在の処理負荷の記録を維持する。

【0017】本発明を理解するために重要なサーバ105乃至107の動作が、それぞれ図3乃至図5の中に示されている。この従来の動作の他に、各サーバは所定の負荷限界値で初期化される負荷制御プログラムを実行する。図3に示されているように、サーバ105は現在の負荷が高負荷上限値「A1' high」を超えているかどうかをステップ300において繰返しチェックする。上限値を超えていなかった場合、それはサーバ105が過負荷になっていないことを意味し、従って、サーバ105はステップ300にとどまっている。負荷が上限値を超えていた場合、それはサーバ105が過負荷であることを意味し、従って、サーバ105は一次ページA1(240)に対して二次ページA1'(241)をステップ302において置き換える。これはデータベースA210の部分AX212からの情報に対するそれ以降のすべての要求がサーバ107に対して向けられるようにする効果がある。サーバ107はこれらの要求を複製部分AX232から従来の方法でサービスする。従って、サーバ105が過負荷になると、その処理負荷のいくつかはサーバ107によって肩代わりされる。

【0018】ステップ302に続いて、サーバ105は高負荷の下限値「A1' low」がその現在の処理負荷を超えているかどうかをステップ304においてチェッ

10

20

30

40

50

クする。超えていた場合、それはサーバ105が過小負荷状態にあることを意味し、従って、サーバ105はステップ306において、二次ページA1' (241) に対して一次ページA1 (240) を置き換える。これは初期動作を回復する効果を持つ。その場合、サーバ105はデータベースA210からの情報に対するすべての要求にサービスしている。次に、サーバ105はステップ300へ戻る。

【0019】限界値A1' lowが現在の負荷を超えていないことがステップ304において判定された場合、サーバ105は現在の負荷が再びA1' highの限界値を超えていないかどうかをステップ308においてチェックする。超えていなかった場合、それはサーバ105が過負荷状態でないことを意味し、従って、サーバ105はステップ304へ戻る。A1' highの限界値を超えていた場合、それはサーバ105が再び負荷になっていることを意味し、従って、サーバ105はステップ310において、二次ページA1' (241) に対して三次ページA1' (242) を置き換える。これはデータベースAの部分AY213からの情報に対するそれ以降のすべての要求がサーバ106に対して向けられるようにする追加の効果がある。サーバ106は複製部分AY223から従来の方法でこれらの要求にサービスし、それによってサーバ105によって行われるはずであった処理負荷のいくつかを肩代わりし、従って、サーバ105の負荷が軽減される。

【0020】ステップ310に続いて、サーバ105はA1' lowの限界値がサーバ105の現在の処理負荷を超えているかどうかを、ステップ312においてチェックする。超えていなかった場合、サーバ105はステップ312の状態にとどまる。超えていた場合、それはサーバ105が過小負荷になっていることを意味し、従って、サーバ105は三次ページA1' (242) に対して二次ページA1' (241) をステップ314において置き換える。これはステップ310においてサーバ106に対して転送された処理負荷の部分をサーバ105が戻す効果がある。次に、サーバ105はステップ304へ戻る。

【0021】サーバ106と107の動作は似ている。図4に示されているように、サーバ106は現在の処理負荷が高負荷の上限値「B3' high」を超えているかどうかを、ステップ400において繰返しチェックする。超えていなかった場合、サーバ106は過負荷状態ではなく、ステップ400にとどまる。超えていた場合、サーバ106は過負荷状態であり、従って、サーバ106は一次ページB3 (252) に対して二次ページB3' (253) をステップ402において置き換える。これはデータベースB220の部分BZ221からの情報に対するそれ以降のすべての要求がサーバ105に対して向けられるようにする効果がある。サーバ10

5はこれらの要求に対して複製部分BZ211から従来の方法でサービスする。それによってサーバ106の負荷が解放される。

【0022】ステップ402に続いて、サーバ106は高負荷の下限値「B3' low」がその現在の処理負荷を超えているかどうかをステップ404においてチェックする。超えていなかった場合、サーバ106はステップ404にとどまる。超えていた場合、それはサーバ106が過小負荷状態であることを意味し、従って、サーバ106は二次ページB3' (253) に対して一次ページB3 (252) をステップ406において置き換える。これは初期の動作を回復する効果があり、サーバ106はデータベースB220からのすべての要求にサービスしている。次にサーバ106はステップ400へ戻る。

【0023】図5に示されているように、サーバ107はその現在の処理負荷が負荷限界値「C1'」を超えているかどうかを、ステップ500において繰返しチェックする。超えていなかった場合、サーバ107は過負荷状態ではなく、ステップ500にとどまる。超えていた場合、サーバ107は過負荷状態であり、従って、それは一次オブジェクトC1 260に対して二次オブジェクトC1' (261) をステップ502において置き換え、それによってその処理負荷の一部をサーバ106へ転送する。ステップ502に続いて、サーバ107は負荷限界値「C1'」がその現在の処理負荷を超えているかどうかをステップ504において繰返しチェックする。超えていなかった場合、サーバ107はステップ504にとどまる。超えていた場合、それはサーバ107が過負荷状態にはなっていないことを意味し、従って、サーバ107はステップ506において、二次オブジェクトC1' (261) に対して一次オブジェクトC1 (260) を置き換える。これによってその初期動作を回復する。次にサーバ107はステップ500へ戻る。

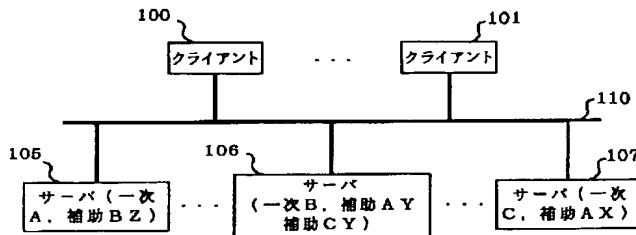
【0024】もちろん、上記の例示としての実施形態に対する各種の変更および修正が可能であることは、この分野の技術に熟達した人にとっては明らかである。例えば、一次および二次のページまたはオブジェクトの両方を格納する代わりに、一次ページまたはオブジェクトが「高速 (on-the-fly)」で (すなわち、リアルタイムで) 二次ページまたはオブジェクトに変換されるようにすること、あるいはその逆が行われるようにすることができる。同様に、データベースの複製部分が補助サーバ上にあらかじめ格納されている代わりに、データベースの部分が「高速」で補助サーバに対して複製されて分配されるようにすることができる。さらに、単位時間当たりのアクセスの回数以外の測定値および限界値を使って、一次サーバとの間の負荷の肩代わりまたは戻しの処理を行うかどうかを判定することができる。これらの測定値および限界値は過去において同様な時刻において経

験された負荷に基づいて、将来の負荷を推定する予測アルゴリズムなどの前方監視 (forward-looking) とすることができる。さらに、メイン・サーバは待機サーバから現在の処理負荷データを要求することができ、そしてこれらのデータをそれらの待機サーバに対する肩代わりの処理を行うかどうかの判定に組み込むことができる。そのような変更および修正は本発明の精神および範囲から逸脱することなしに、そしてその利点を減らすことなしに行うことができる。従って、そのような変更および修正は特許請求の範囲によってカバーされることが意図されている。

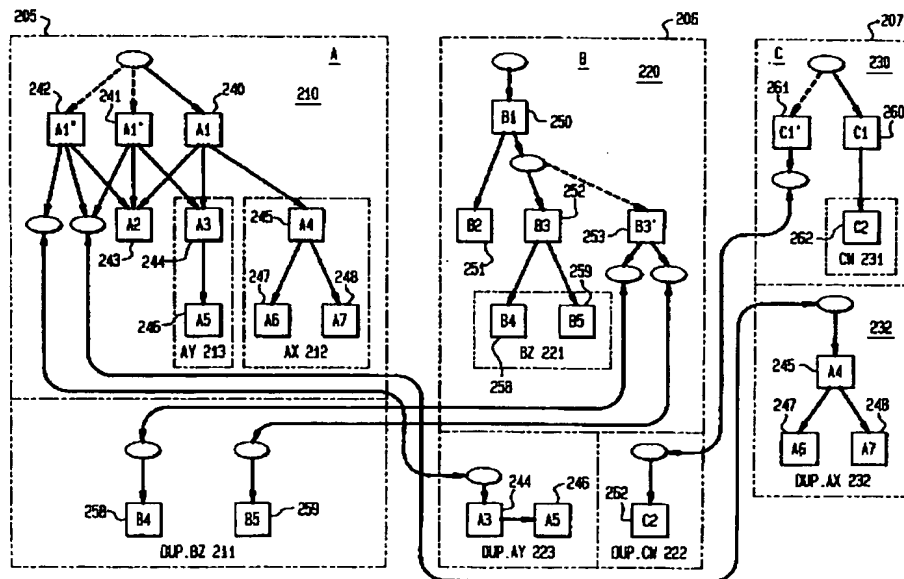
【図面の簡単な説明】

【図1】 本発明の例示としての実施形態を含む情報ネッ

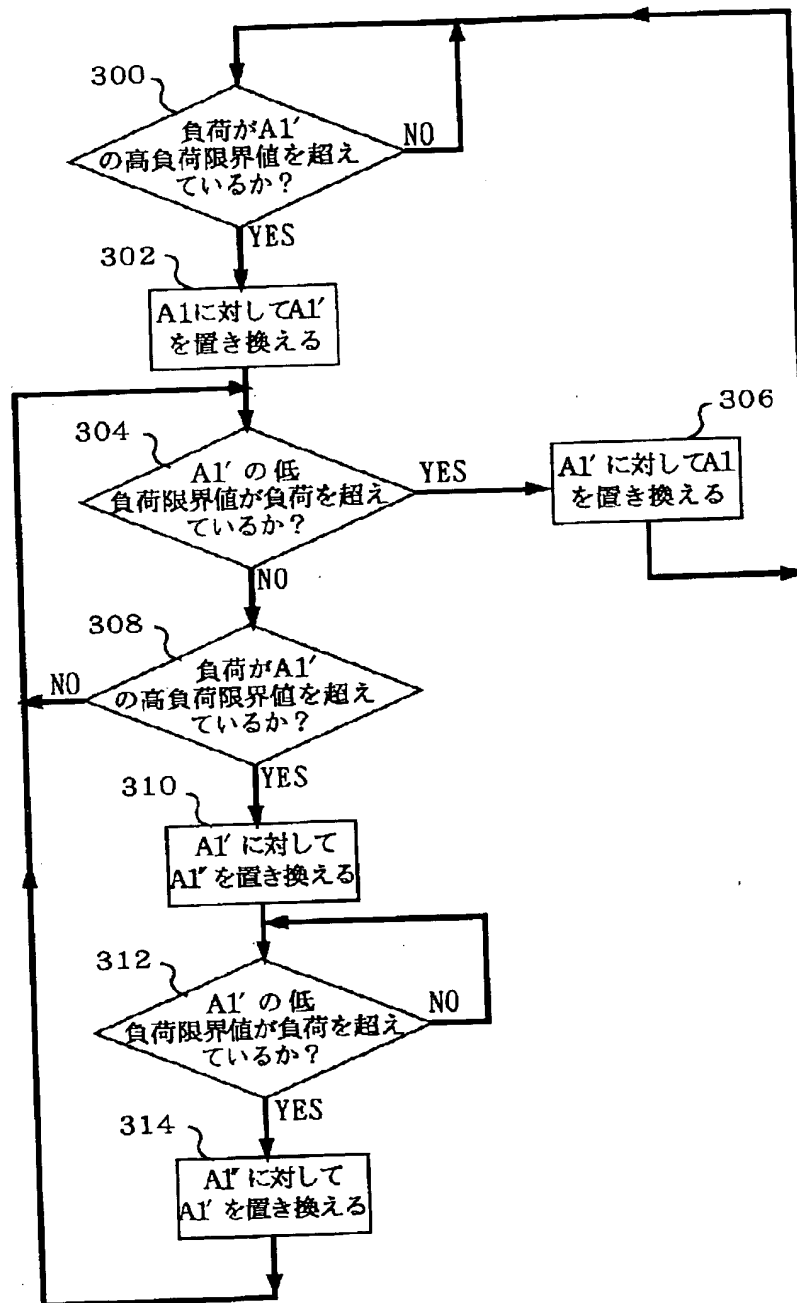
【図1】



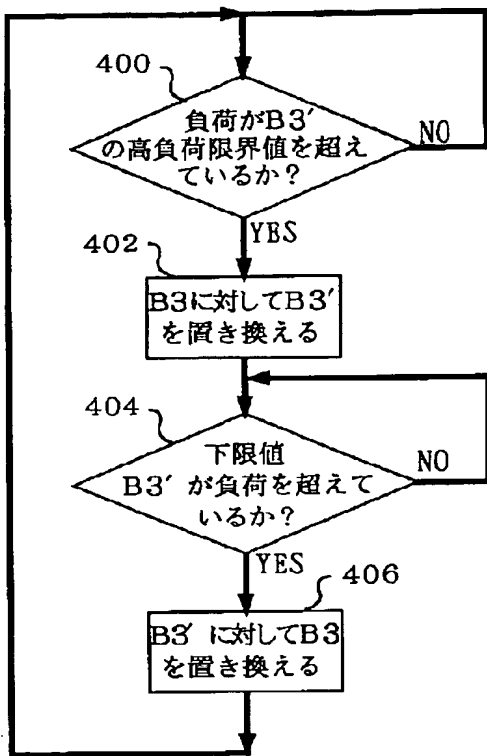
【図2】



【図3】



【図4】



【図5】

